Т.В. Дзибук (Минск, МГЛУ)

ЛИНГВИСТИЧЕСКАЯ БАЗА ДАННЫХ СИСТЕМЫ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ФАКТОВ ИЗ ТЕКСТА АВТОБИОГРАФИИ И РЕЗЮМЕ

По мнению многих специалистов в области компьютерных телекоммуникационных технологий массивы неструктурированных данных (текстов) составляют большую часть информации, с которой имеют дело пользователи. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта специализированные технологии типа Information Extraction, Data- и Text Mining осуществляют смысловой анализ и поиск различных фактов, понятий, описаний информационных объектов и т.п. в неструктурированных данных.

Моделирование процесса автоматического извлечения фактов из текстов досье — автобиографий и резюме проводилось на материале 40 англоязычных текстов указанного типа, взятых со страниц различных сайтов сети Интернет. Разработанная формальная модель системы базируется на лингвистической базе данных. Прежде чем рассмотреть процедуру ее создания, необходимо отметить следующее. Несмотря на то, что изучаемые тексты относятся к текстам официально-делового стиля, предполагающего стандартизацию лексических единиц и структуры оформления, при выделении некоторых текстовых элементов необходимо учитывать описанные ниже особенности.

Автобиография. Имя автора автобиографии содержится в первом предложении текста после маркера *name*, т.е. маркер служат левой границей искомого факта. Правой границей является знак препинания. Количество слов, составляющих имя, не является фиксированным. В ряде документов указан возраст автора (с использованием букв или цифр). Однако, принимая во внимание подверженность данного параметра постоянному изменению, для более долгосрочного использования базы данных с извлеченными фактами необходимо находить (по возможности) дату рождения. Кроме того, если указана дата рождения, возраст может быть определен с учетом текущей даты. Для описания дат используется ряд шаблонов, выбор варианта/вариантов которых является прерогативой автора текста. Среди вариантов представления дат были выделены следующие: *in 1961; in January 1961; 12/05/1961; on the 5th of January, 1961; on the fifth of January, 1961*. В большинстве исследованных текстов, в случаях, когда информация о событии содержит и дату, и место события, дата указана перед обозначением места.

При выделении адреса проживания необходимо учитывать, что данный параметр мог меняться на протяжении жизни автора. Маркером в таких ситуациях служит глагол *moved to* с указанием нового адреса проживания. Следует отметить, что наличие в тексте документа места рождения и отсутствие информации о текущем месте жительства автора не является

основанием того, что автор по-прежнему проживает по месту рождения. Возможно, информация была опущена в связи со спешкой, фиксацией данных в других документах (например, в прилагающемся к автобиографии резюме) и отсутствием необходимости дублирования, по личным мотивам и т.д.

Для поиска информации об учебной деятельности автора могут использоваться слова, обозначающие виды учебных заведений, например, *school, college, university*. Дополнительные характеристики отмечены такими единицами, как *elementory, junior, senior*, а также уникальными наименованиями учебных заведений (если учебное заведение названо в честь известного деятеля искусства, политики и т.д.). Если в учебном заведении автор текста автобиографии познакомился с будущей/будущим супругой/супругом, в предложениях об учебном заведении может встречаться местоимение *we* в таких контекстах, как *We graduated, we continued our education*. В связи с этим, при поиске фактов целесообразно обращать внимание на глаголы и игнорировать местоимения. Если автор в данный момент является студентом, период обучения указан следующим образом: *год поступления — present*.

Для описания профессиональной деятельности используется большое количество шаблонов, содержащих следующие элементы: должность, место работы, период работы (стаж). Как правило, данные элементы расположены в документе в аналогичной последовательности. Левым маркером стажа работы служит предлог *for*, имеющий в подобном контексте временное значение. Степень детализации излагаемой информации определяется автором. Так, в ряде образцов за названием компании следует ее краткое описание (*a company [name of the company] which is…; a company [name of the company] that …* (описание компании)). С другой стороны, слова *which* и *that* в таком контексте может служить правым маркером при выделении названия компании.

Фрагменты с краткой информацией о близких родственниках содержат ключевые слова, обозначающие степень родства. Семейное положение автора описано указанными в таблице выражениями.

Резюме. Одним из отличий текста резюме от текста автобиографии является представление информации в виде блоков, что позволяет активно использовать эти структурные особенности компьютерной системой. В большинстве исследованных текстов англоязычных резюме блоки имеют заглавия, что облегчает структуризацию документа и позволяет сузить область поиска необходимых фактов.

Следует отметить, что, в зависимости от специфики должности, на которую претендует соискатель, текст резюме может содержать дополнительные разделы, наличие и содержание которых невозможно предугадать заранее. В связи с этим, в базе данных находятся только те элементы, которые были выделены в подавляющем большинстве документов. Так, имя и фамилия автора текста указаны в заглавной части документа (первая строка), часто без маркера *Name*. Встречаются резюме, в которых фамилия

и имя дополнительно указаны в колонтитулах, однако для автоматического извлечения данных фактов достаточно проанализировать верхнюю часть документа. Дата и место рождения, являясь важными компонентами текста автобиографии, в то же время не указываются в текстах резюме, поскольку важным фактом для приема на работу является текущее место жительства.

В верхней части текста документа указана также контактная информация: номера телефонов, адреса электронных почтовых ящиков. Контактные данные могут сопровождаться пометками *Home, Office, Mobile, Cell.* Номера телефонов могут быть опознаны по последовательности цифр, электронные адреса — по обязательному символу @. Следует отметить, что в автобиографиях контактная информация отсутствует.

Специфическим для резюме является также блок (строка) *Objective*, содержащий должность, на которую претендует соискатель. Информация об образовании находится в блоке, который может быть озаглавлен следующим образом: *Education, Education and further training*. В резюме могут быть указаны полученные за время учебы сертификаты и отметки по дисциплинам (раздел *Subjects and grades*), а также дополнительно пройденные курсы (с пометкой *Additional Courses*). Информация о временных рамках содержится в правой или левой части документа. Часто указывается только год окончания обучения. Название блока, содержащего информацию о профессиональной деятельности, может варьироваться: *Experience, Professional Experience, Employment History. Work Experience. Construction Career Summary, Relevant Experience.* В одной строке с наименованием специальности содержится период работы. Дополнительно может присутствовать пометка *Part time*, свидетельствующая о частичной занятости на должности. Под наименованием должности (следующая строка/строки) указано место работы, которое содержит наименование организации и в некоторых случаях — ее адрес. Кроме того, для каждой специальности приведен перечень исполняемых обязанностей. Информация такого рода не заносится в базу данных, так как не является выделяемым элементом. Стажировки отмечены маркером *Intern*. Информация о профессиональных качествах содержится в блоках *Skills, Key skills, Computer skills*. Последний блок содержит перечисление программных продуктов, с которыми может работать соискатель (*MS Excel, Word, Publisher*), и уровень владения указанными продуктами и навыками (*proficient in…; advanced level; word processing (typing) N wpm (words per minute) N % accuracy)*. Уровень владения иностранными языками осуществляется в форме *language : level*. Личные качества могут быть зафиксированы в разделе *Personal Qualities*.

В зависимости от целей соискателя, в резюме могут встречаться и нестандартные элементы. Так, в одном из исследованных текстов встретился раздел *Personal Data* со следующими параметрами: *height, weight, hair, eyes*. В ряде образцов имеется блок *Referees* (имена и контактные данные лиц, готовых обеспечить заинтересованную сторону информацией о профессиональной деятельности соискателя).

Фрагмент базы данных представлен ниже в таблице 1.

Таблица

Факт	Опорное слово	Шаблон (!...! — реализация факта)
Фамилия, имя	*Name*	My name is !...! I am !...! by !...! (в заголовке)…
Дата рождения	*Born*	I was born in/on !...! My birthday is on !...!... варианты представления дат: [year] [month] [year]…
Место рождения	*Born*	I was born in !...!
Место жительства	*From, live, grew up, moved to*	I'm from !...! (city, country) We live (together) in !...! Moved to !...! …
Возраст	*Years old*	(now) I am !...! years old
Родственники	*family, mother, father,* *sister, brother, child,* *children etc.* *adopted at birth*	My mother/father/sister/brother is (a) !...! My mother/father/sister/brother's name is !...! There are !...! of us in our family (number of family members) There are !...! members in our family Our family has !...! members I have in my family !...! I have N children…
Семейное положение	*married, not married,* *got married* *bachelor*	I am !...!
Образование		
– место учебы	*university, college, school (named),* *elementary school,* *junior high school* *senior high school*	I am a/an (international) student at !...! I entered !...! in (year) I went to !...! I studied in !...! for (time) In (year) I began attending !...! The school's name is !...! I moved to (city) and began school there…
– период учебы	*Went, finished, entered*	I studied in/at… from !...! to !...! In !...! I graduated from !...! I entered (place) in !...! …
– специальность	*major, degree*	I studied to be !...!; I majored in !...!; I studied !...! at (place)…
–научная степень	*Degree*	I hold a !bachelor's! degree with a major in !...!
Профессиональная деятельность		
– место работы	*Position, (private) company, firm, office*	After I finished my course, I started to work at !...! I worked as (occupation) for !...! I worked in the !...! !(alternative name of the company)! ….

Продолжение табл.

Факт	**Опорное слово**	**Шаблон (!...! — реализация факта)**
– должность	*Work(ed), job, part-time job, position*	work/worked(…)as (a)!...! <…> became a !...! I worked as !...! for (place)
– период работы	*month(s), year(s) retired, quit*	I worked (there)…for !...! months/years I quit the company in !...! (to set up my own business)
Служба в армии	*army* *military service* *retired* *served*	joined the army military service retired from the army served in the military for N years

Перечисленные выше маркеры, указывающие на какой-либо факт в исследованных текстах автобиографий и резюме, позволили выделить определенные шаблоны, которые и составили лингвистическую базу данных системы автоматического извлечения фактов из текстов данного типа.